

本周周报(1.7-1.13):

解聪

本周工作:

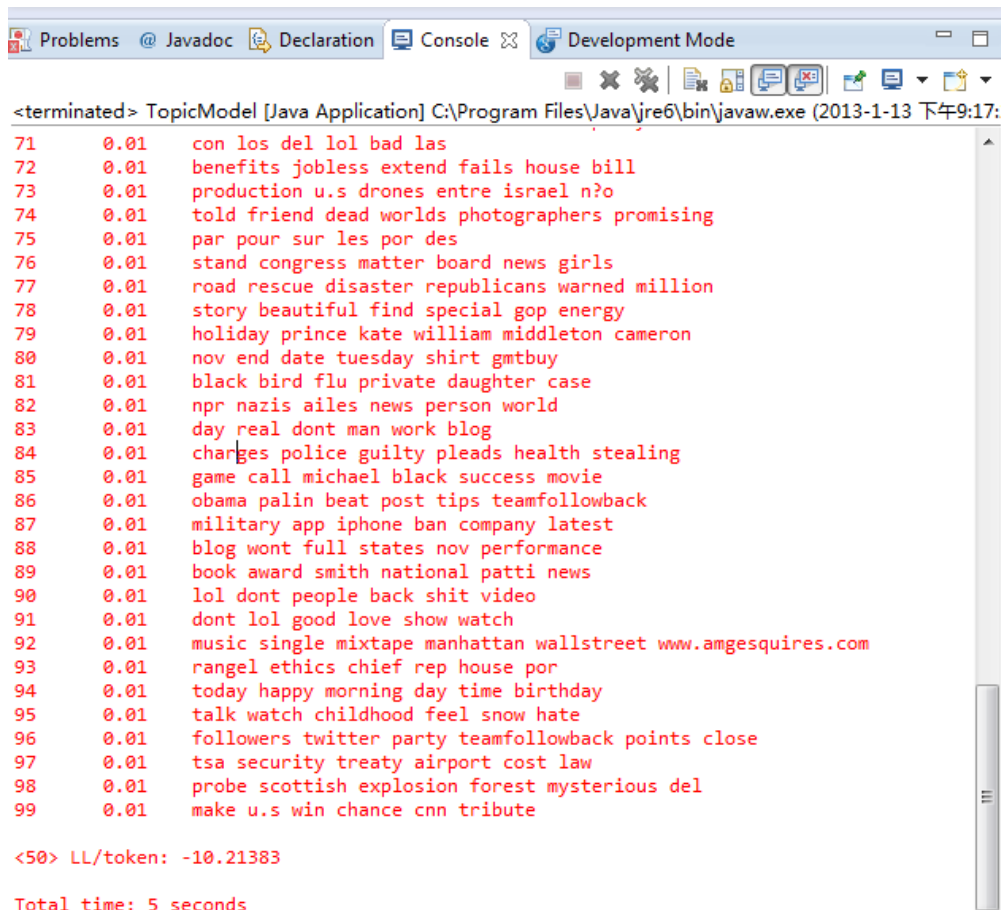
1. 时序用户数据分析:

本周主要继续对案例分析进行改进以及数据的改进。

案例分析的改进

Twitter 数据分析本周进行了改进。主要是采用了 LDA 对 twitter 的数据进行了主题分析，算法主要采用了 MALLET 包中现有的主题建模的实现。

先前的可视化实现文本分析这部分主要是由手动完成的。因此，先前的方法就直接选取了一个较小的子集（仅包含 600 条 twitter）以减小手工操作的工作量。现在就直接对一天的 3 万多条数据进行分析。由于 LDA 中需要指定抽取的主题数，类似于 k-means 的 k 值。所以初步尝试了一下。选择的依据是根据每条 Twitter 最后对应到主题的贴切程度。如果主题数很多，结果比较细，但是不够概括，而且程序效率低。主题数较低，Twitter 对应到的主题不够准确。在初步尝试以后，发现大概可以将每天的主题数定在 100 左右。



```
<terminated> TopicModel [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (2013-1-13 下午9:17:
71      0.01      con los del lol bad las
72      0.01      benefits jobless extend fails house bill
73      0.01      production u.s drones entre israel n?o
74      0.01      told friend dead worlds photographers promising
75      0.01      par pour sur les por des
76      0.01      stand congress matter board news girls
77      0.01      road rescue disaster republicans warned million
78      0.01      story beautiful find special gop energy
79      0.01      holiday prince kate william middleton cameron
80      0.01      nov end date tuesday shirt gmtbuy
81      0.01      black bird flu private daughter case
82      0.01      npr nazis ailes news person world
83      0.01      day real dont man work blog
84      0.01      charges police guilty pleads health stealing
85      0.01      game call michael black success movie
86      0.01      obama palin beat post tips teamfollowback
87      0.01      military app iphone ban company latest
88      0.01      blog wont full states nov performance
89      0.01      book award smith national patti news
90      0.01      lol dont people back shit video
91      0.01      dont lol good love show watch
92      0.01      music single mixtape manhattan wallstreet www.amgesquires.com
93      0.01      rangel ethics chief rep house por
94      0.01      today happy morning day time birthday
95      0.01      talk watch childhood feel snow hate
96      0.01      followers twitter party teamfollowback points close
97      0.01      tsa security treaty airport cost law
98      0.01      probe scottish explosion forest mysterious del
99      0.01      make u.s win chance cnn tribute

<50> LL/token: -10.21383

Total time: 5 seconds
```

程序运行 LDA 的部分结果，每行表示每个主题对应的一些关键词，主题个数为 100。文本量为一天的 3 万多条微博。

可以发现有些很奇怪的主题，比如 71 行的主题包含的关键词。这些虽然使用的是英文字符，但是他们都不是英语而是菲律宾语。一方面，会对原有的英文主题的抽取产生影响；

另一方面，虽然 LDA 是基于词频分析的，对于类似于这种以单词为单位的语言仍然有效，但是我们并没有滤掉这些语言中的常见词(对于英语来讲就是 the for 等词)，所以其实即便抽取了效果也不是很好。因此 Twitter 数据还需有所过滤或重新抓取。

下周会做 twitter 数据中自然语言处理对情感分析的部分工作。

淘宝数据分析本周进展不大，总体方向是寻找对每个时刻的多个高维变量的代表，即全局的指标。比如，对于 Twitter 数据我们可以使用主题抽取来表示每个时刻的多个高维变量以及其总体演化。

总结

无论对于淘宝，还是对于 twitter 数据，我觉得可以统一到同一个分析流程中。而且对其分析的流程分为两个层次：

层次	分析方法	分析对象	特定数据	对应含义	分析方法	实现情况	数据现状
全局层面	需要部分数据分析算法的辅助	反映所有时序记录的整体趋势	Twitter: 主题演化	包括突发的事件, 如美国总统大选等。	有方法: LDA+STL	已实现	有数据, 不完善
			淘宝: 还不太清楚, 但是也可以反应交易的演化	比如诺贝尔奖前后莫言作品的大量交易。	尚未明确	未实现	有数据, 不完善 (没有特定日期的数据)
局部层面	可以通过可视化直接完成	针对特定数据, 往往和特定的时间点与特定的频率有关。	淘宝: 就是个别交易模式	正常模式: 商家促销	现有方法: 可视编码设计 + 部分数据分析	已实现, 需改进 (可能含有异常的时间段分析的实现)	有数据
				异常模式: 买家刷信誉			
			Twitter: 个别转发模式	正常模式: 名人效应	现有方法: 可视编码设计 + 部分数据分析	已实现, 需改进 (Twitter 数据情感分析的实现)	无数据
				异常模式: Twitter 水军			

注：红色到绿色表示完成程度由低到高。

因此，可以发现，其实淘宝数据和 Twitter 数据的模式极为相似。完全可以统一到一个泛化的分析流程。问题在于 twitter 数据是大家经常分析的数据，分析方法较为成熟，而淘宝数据则还不太清楚如何处理。

按照上面的表格可以发现其实现在完成的还是属于一小部分。为实现的部分还包括算法框架对各种数据的整合。

2. Twitter 数据的抓取

Twitter 数据不够完整。目前的数据是 1600 多人的用户群的 Twitter 信息，解决问题的方法是重新抓一些有用的，可以反应时序模式的数据。本周使用 `twitter4j` 对数据进行抓取，但是返回 404 的错误。经分析，可能是因为国内无法登陆 `twitter` 的原因。使用代理服务器之后可以顺利登陆并获得数据，这部分工作预计下周可以实现。

3. 论文书写

现有的论文的上传到 `svn`。目前前两章使用英文，后面几张使用中文（因为方法还不完整）。中文做成 PDF 不会显示。

地址：`svn://zjuvag.3322.org/Projects/VIS2013/UserBehavior`

4. Taobao 的其他工作

帮助数据魔方定制可视化形式。

下周工作：

1. Twitter 数据的抓取，Twitter 文本的情感分析。
2. 时序用户行为可视化英文文档书写